The Need for Truly Graded Lexical Complexity Prediction

David Alfter
Gothenburg Research Infrastructure in Digital Humanities (GRIDH)
University of Gothenburg
Sweden
david.alfter@gu.se

Abstract

Recent trends in NLP have shifted towards modeling lexical complexity as a continuous value, but practical implementations often remain binary. This opinion piece argues for the importance of truly graded lexical complexity prediction, particularly in language learning. We examine the evolution of lexical complexity modeling, highlighting the "data bottleneck" as a key obstacle. Overcoming this challenge can lead to significant benefits, such as enhanced personalization in language learning and improved text simplification. We call for a concerted effort from the research community to create high-quality, graded complexity datasets and to develop methods that fully leverage continuous complexity modeling, while addressing ethical considerations. By fully embracing the continuous nature of lexical complexity, we can develop more effective, inclusive, and personalized language technologies.

1 Introduction

Lexical complexity prediction (LCP) is the task of assigning a complexity score to a word or phrase, indicating how difficult it is to understand for a given target population, such as language learners or readers with disabilities (Shardlow et al., 2022). In recent years, the field of Natural Language Processing (NLP) has witnessed a shift in approach to lexical complexity prediction. There has been a growing recognition that lexical complexity is not a binary concept, but rather exists on a continuum (Shardlow et al., 2020). This acknowledgment has led to efforts to model lexical complexity as a continuous value, promising more nuanced and accurate representations of word difficulty across various contexts and for different readers.

However, despite this conceptual advancement, the practical implementation of truly graded lexical complexity prediction remains limited. This discrepancy between theoretical understanding and applied research is evident in recent shared tasks and datasets in the field. While the 2021 SemEval shared task on LCP (Shardlow et al., 2021) made strides by including multiple contexts for about half of its training instances, subsequent initiatives have not fully embraced this approach. Notably, the 2024 MLSP (Multilingual Lexical Simplification and Prediction) task (Shardlow et al., 2024) included only a single word with two contexts, effectively reverting to a predominantly one-to-one mapping of words and complexities.[1]

This persistent focus on one-to-one complexity mapping not only fails to capture the full spectrum of lexical difficulty but also hinders progress in areas where truly graded predictions are crucial. One such domain is language learning, where learners progress through various levels of proficiency and require finely-tuned assessments of word difficulty (Crossley et al., 2017; Gooding et al., 2021). In this context, binary classifications of "simple" or "complex" are insufficient to guide effective vocabulary acquisition strategies or to develop adaptive learning materials. Further, polysemous words generally show a spread of word senses over different levels, and not all meanings are learned or known at each level (Alfter et al., 2022).

This opinion piece argues that the field of NLP must move beyond its current limited implementation of continuous lexical complexity modeling. We contend that embracing truly graded predictions is not just a matter of theoretical correctness, but a necessity for advancing practical applications in areas such as language learning, text simplification, and readability assessment. By doing so, we can develop more sophisticated and useful tools that accurately reflect the nuanced nature of lexical complexity across diverse contexts and user needs.

Consider the word "crane". In the context of

[1] The test set ($n = 5123$), which may be used as additional training data after the completion of the task, contains 4% of sentences for a word with more than one context.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 326–333
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

construction, “crane” refers to a machine used for lifting and moving heavy objects, which might be a familiar concept to most adult readers. However, in the context of ornithology, “crane” refers to a family of large, long-legged birds, which might be less familiar to readers without a background in bird watching or biology. A genuinely graded lexical complexity prediction system should be able to assign different complexity scores to “crane” based on its context, reflecting the varying levels of difficulty for different readers.

A crucial dimension currently underrepresented in lexical complexity research is an explicit theoretical analysis of the construct itself. Lexical complexity is inherently multidimensional, encompassing orthographic difficulty (Just and Carpenter, 1987; Perfetti et al., 2005; Alfter, 2021), conceptual complexity (Nation and Nation, 2001), atypical contextual usage (Erk and Padó, 2008; Peters et al., 2019), and figurative or metaphorical meanings (Steen et al., 2010; Thibodeau and Boroditsky, 2011). These distinct aspects significantly impact different user groups in varied ways; for example, native children encountering conceptual complexity differ from adult second-language learners struggling primarily with orthographic unfamiliarity or contextual atypicality (Akamatsu, 2005; Crossley and McNamara, 2012).

2 Current State of the Field

The field of lexical complexity prediction and simplification has evolved significantly over the past decade, with researchers exploring various approaches to model and predict word difficulty. This section provides an overview of key developments and current trends in the field.

2.1 The Divide Between Two Worlds

In this section, we highlight two related yet disconnected main fields active in lexical complexity prediction: lexical complexity prediction for lexical simplification, and lexical complexity prediction for language learning applications.

Lexical simplification can have a broad range of applications, most aiming at making texts easier to read for certain audiences such as children (De Belder et al., 2010), language learners (Petersen and Ostendorf, 2007; Rets and Rogaten, 2021), people with reading disabilities (Devlin, 1998; Chung et al., 2013), simplifying medical texts (Deléger and Zweigenbaum, 2009) or judicial texts (LoPucki, 2014), to name but a few. In this line of research, an important first step is to identify *complex* words (Specia et al., 2012). This line of research in lexical complexity prediction started as *complex word identification* (Shardlow, 2013), a binary classification tasks of words into *simple* and *complex* words. Shardlow (2013) presented one of the first comprehensive studies on automatic lexical simplification, focusing on identifying complex words and suggesting simpler alternatives. This binary approach was further developed in subsequent studies, such as Paetzold and Specia (2016b), who introduced a feature-based machine learning approach to complex word identification.

At around the same time, another line of research emerged: graded lexical complexity prediction (Gala et al., 2013, 2014). The main difference to complex word identification is that the aim is to predict a *grade* for each word, corresponding to different school levels for native language learners, and later second language learner proficiency levels (Tack et al., 2016; Alfter et al., 2016; Alfter and Volodina, 2018b; Tack et al., 2018; Pintard and François, 2020). This line of research is tightly connected to (second) language acquisition, with applications such as adaptive learning content (Burstein et al., 2017; Alfter and Graën, 2019) and personalized models for vocabulary learning (Avdiu et al., 2019; Ehara et al., 2018; Yancey and Lepage, 2018).

Over time, the two fields moved closer together, with complex word identification becoming *lexical complexity prediction*, with the aim of predicting a continuous complexity value instead of binary labels. Despite this, it remains that LCP for lexical simplification is concerned with finding words that should be simplified, while LCP for language learning purposes is concerned with finding words that are suitable for learners of a given proficiency level.

2.2 Shared Tasks and Datasets

Shared tasks have played a crucial role in advancing the field. In 2016, the first Shared Task on Complex Word Identification (Paetzold and Specia, 2016a) was organized, followed by the 2018 CWI Shared Task on Complex Word Identification (Yimam et al., 2018). In 2016, the data targeted only English, while in 2018, the task introduced multilingual and cross-lingual complex word identification, but still treating the problem as binary.[2] A significant shift occurred with the 2021 SemEval shared task on Lexical Complexity Prediction (Shardlow et al., 2021), which introduced a dataset with continuous complexity scores derived from Likert scale annotations and multiple contexts for many words. This task represented a major step towards more nuanced modeling of lexical complexity.

Despite the progress towards continuous modeling, recent work still shows a tendency to simplify the problem. The 2024 MLSP task (Shardlow et al., 2024), while advancing the multilingual aspect, largely reverted to a one-to-one mapping with limited contextual variation. The training data ($n = 300$) contains a single word with exactly two different contexts and almost identical complexity values. We argue that this is egregiously insufficient to learn different complexities for the same word in different contexts. This setup effectively reduced the task to a one-to-one mapping of words and complexities, disregarding the context-dependent nature of lexical complexity that was captured in the CompLex dataset. In opposition, the 2021 shared task training data ($n = 3487$) contains 1701 words with multiple contexts and different complexity values.

2.3 The Problem

Ideally, one would want to capture context-specific complexity and train systems to automatically predict such complexity. In order to train a system to recognize context-specific complexity, or *truly* graded complexity, the training data would have to include multiple contexts per word with *varying* complexity values. Even though complex word identification moved towards continuous modeling of complexity, it still often only gives one context per word, effectively mapping one word to one complexity value.

Recent research shows that out-of-the-box large language models are not capable of efficiently grading vocabulary (Alfter, 2024; Kelious et al., 2024). This at least to some degree precludes the use of large language models for synthetic data creation. If one were to for example build a system to automatically generate proficiency-adapted definitions, one would need to fine-tune a model with truly graded data (Yuan et al., 2022).

3 Data Bottleneck

While theoretical advancements in lexical complexity prediction have pushed towards more nuanced, continuous modeling, a significant obstacle impedes practical implementation: the data bottleneck. This section explores the challenges in obtaining and creating the rich, context-aware datasets necessary for truly graded lexical complexity prediction.

3.1 Data Scarcity

The shift from binary to continuous lexical complexity modeling demands datasets that capture fine-grained distinctions in word difficulty. However, such resources are rarely available at the scale required for robust model training. As noted by Shardlow et al. (2022), creating datasets with continuous complexity ratings is significantly more resource-intensive than binary labeling tasks. Their study found that annotators spent an average of 21.61 seconds per annotation for graded complexity ratings.

The CompLex dataset (Shardlow et al., 2020) represented a step forward by providing continuous complexity scores, but even this resource was limited in size and scope compared to larger binary datasets. CompLex contained 10,800 instances across three genres, which, while substantial, pales in comparison to binary datasets like the one used in the 2018 CWI Shared Task, which contained over 65,000 instances (Yimam et al., 2018).

3.2 Challenges in Dataset Creation

Several factors contribute to the difficulty in conceiving and creating appropriate datasets for graded lexical complexity prediction. One significant challenge lies in the subjective nature of assigning precise, continuous complexity scores to words in context. This task demands skilled annotators yet often leads to low inter-annotator agreement (North et al., 2023), although attempts a mitigating this issue have been made using comparative judgments (Gooding et al., 2019; Alfter et al., 2021, 2022).

Another obstacle is the contextual variation inherent in language. Capturing the full spectrum of contextual variations for each word exponentially increases the annotation effort. The 2024 MLSP task's inclusion of only one word with two contexts illustrates the practical challenges in scaling contextual annotations.

Furthermore, considerations regarding annotator

[2] The task consisted of two subtasks, binary and continuous prediction. However, the continuous labels were obtained by averaging the binary labels over all annotations. We thus regard this task as mainly binary.

characteristics such as linguistic background and language proficiency further complicate dataset creation. Differences between native speakers, teachers, and language learners with varying language proficiency levels can lead to significant variations in perceived lexical complexity, thus limiting the comparability and interpretability of the data. Therefore, a clear definition and control of annotator demographics is essential to ensure the validity and usefulness of complexity-annotated corpora.

In addition, lexical complexity can vary significantly across domains, genres and tasks (e.g., reading aloud, reading for comprehension). Creating datasets that adequately represent this diversity while maintaining consistent annotation quality is a formidable task.

Moreover, complexities introduced by figurative language, including metaphors and metonymies, pose challenges, as such uses often deviate substantially from literal meanings, complicating complexity assessment. Similarly, multi-word expressions (MWEs) introduce unique difficulties because their complexity cannot be straightforwardly derived from the complexity of their constituent words (Alfter and Volodina, 2018a).

Finally, extending graded complexity prediction to multiple languages compounds the resource scarcity. Multilingual datasets like the one used in the 2018 CWI Shared Task are rare and often revert to simpler, binary annotations to maintain feasibility across languages.

3.3 Impact on Model Development and Evaluation

The data bottleneck has cascading effects on the field. Without access to large-scale, graded complexity datasets, researchers often default to simpler binary models or resort to synthetic data generation, potentially limiting model sophistication and real-world applicability. Large-scale extensive annotated datasets allow for more comprehensive coverage of phenomena such as ambiguous words, figurative language use, and multi-part expressions that may be inadequately represented in smaller datasets. Furthermore, larger datasets increase model sensitivity to subtle contextual variations, reduce bias, and improve prediction accuracy in diverse linguistic contexts.

The scarcity of diverse, graded datasets also makes it difficult to comprehensively evaluate models' performance across different contexts, domains, and languages. This can lead to overfitting to specific datasets and poor generalization. Additionally, the relative abundance of binary complexity datasets inadvertently reinforces the continued use of binary approaches, creating a cycle that slows the adoption of truly graded prediction methods.

4 Addressing the Data Bottleneck

To move towards truly graded lexical complexity prediction, it is crucial to develop strategies for creating large-scale, diverse datasets that capture the context-dependent nature of word complexity. In this section, we propose several approaches that could help overcome the data bottleneck.

4.1 Collaborative Annotation Efforts

One approach to creating larger, more diverse datasets is to foster collaborative annotation efforts within the research community. By pooling resources and expertise, one can develop shared annotation guidelines and distribute the workload across multiple institutions. This collaborative approach has been successful in other NLP tasks, such as the creation of the Universal Dependencies treebanks (Nivre et al., 2016). Establishing a similar initiative for lexical complexity annotation could help accelerate the development of high-quality datasets. Further shared tasks on the subject may also help.

4.2 Crowdsourcing and Human Computation

Crowdsourcing platforms, such as Amazon Mechanical Turk, have been used extensively in NLP for data collection and annotation (Snow et al., 2008). By leveraging the power of human computation, one can gather lexical complexity annotations from a diverse pool of participants, potentially covering a wider range of contexts and reader backgrounds. This approach has been successfully used to annotate the CompLex data (Shardlow et al., 2022), to gather *comparative* judgments on lexical difficulty (Alfter et al., 2021, 2022), and to collect age-of-acquisition data (Kuperman et al., 2012; Green et al., 2025). However, quality control mechanisms must be put in place to ensure the reliability of crowdsourced annotations (Sheng et al., 2008).

4.3 Mining Graded Textbook Corpora for Lexical Complexity

Graded textbook corpora, which consist of textbooks designed for language learners at different proficiency levels, offer a promising resource for creating lexical complexity datasets. These textbooks are carefully crafted to introduce vocabulary and grammatical structures in a gradual, level-appropriate manner, making them a valuable source of information about word complexity in context.

Graded textbook corpora can be leveraged to derive lexical complexity scores by aligning the vocabulary in each level with language proficiency frameworks like CEFR (Council of Europe, 2001). The relative difficulty of words can be determined by analyzing their distribution across proficiency levels. Words frequently appearing in beginner-level textbooks but rarely in advanced ones would receive lower complexity scores compared to those introduced at higher levels. This approach has been explored in the English Vocabulary Profile (Capel, 2010) and the CEFRLex project[3], which created CEFR-aligned vocabulary lists from graded textbook corpora.

To extend this approach to lexical complexity prediction, one could leverage techniques from natural language processing, such as word embedding models (Mikolov and Dean, 2013) and contextual language models (Devlin et al., 2018), to capture the semantic and syntactic properties of words in context. By combining these models with the complexity information derived from graded textbook corpora, it may be possible to develop more accurate and context-aware lexical complexity prediction systems.

4.4 Leveraging Large Language Models

Recent advances in large language models offer an attractive avenue for addressing the shortage of richly annotated lexical complexity data. By leveraging large language models (LLMs), researchers can systematically generate diverse contexts for vocabulary items, varying key factors such as linguistic complexity, domain specificity, or target proficiency levels (Alfter, 2024; Kelious et al., 2024). Such synthetic data creation methods could transform even simple, context-free word lists into extensive datasets (Yuan et al., 2022; Green et al., 2025). However, the reliability of LLM-generated complexity annotations would require careful validation, as the generated contexts and associated complexity levels may not align accurately with intended proficiency targets, necessitating subsequent human verification or iterative refinement processes.

5 Conclusion

As we have explored throughout this opinion piece, the field of lexical complexity prediction stands at a critical juncture. While recent trends have acknowledged the continuous nature of word difficulty, practical implementations largely remain tethered to binary and one-to-one paradigms. This disconnect between theoretical understanding and applied research impedes progress in areas where truly graded predictions are not just beneficial, but essential.

The persistence of binary and one-to-one mapping methods is not due to a lack of theoretical understanding, but rather stems from a critical data bottleneck. Creating rich, context-aware datasets with continuous complexity ratings is a formidable challenge, requiring significant resources and expertise. This scarcity of nuanced data has cascading effects, limiting model sophistication and evaluation, and inadvertently reinforcing simpler binary paradigms.

Careful consideration of potential user groups is essential to effectively guide the creation and evaluation of lexical complexity datasets. While our discussion primarily focused on second language learners, graded lexical complexity is also suitable for native speakers and various user contexts, such as readability assessments, literacy support, text accessibility, and the development of educational resources. Each user group may require different complexity scales (e.g., continuous numerical scales suitable for NLP systems to discrete scales aligned with educational frameworks such as CEFR proficiency levels or school grades). Future research should explicitly explore these diverse user needs, considering practical implications such as scale granularity and annotation methods to ensure that lexical complexity annotations are both practically relevant and broadly applicable.

In conclusion, the future of lexical complexity prediction lies in not only fully embracing its continuous nature but also in creating resources that reflect various complexity values per word, allowing for the training of *truly* graded lexical complexity prediction systems. By moving beyond binary simplifications and overcoming the data bottleneck, we can develop tools and applications that more accurately reflect the nuanced reality of language complexity.

[3] https://cental.uclouvain.be/cefrlex/

Limitations

As this is an opinion piece, our focus has been on identifying theoretical limitations and potential avenues for future research within the field of computational lexical complexity modeling. We have not conducted empirical experiments or proposed specific algorithms or datasets. Instead, we have highlighted general shortcomings in existing data and methods and suggested potential directions for advancement.

For the sake of conciseness, we focus on two areas only, namely lexical simplification and language learning. We acknowledge that the implications may reach further than just these two fields.

Ethical Concerns

While the potential benefits are significant, implementing truly graded lexical complexity prediction also presents challenges and ethical considerations. Complexity predictions must account for cultural and linguistic diversity to avoid perpetuating biases. What is considered complex in one cultural or linguistic context may not be in another.

The detailed learner data required for personalized systems raises privacy concerns. Ethical guidelines for data collection and use in educational technology must be carefully considered.

Acknowledgements

This work was financially supported by the Royal Society of Arts and Sciences in Gothenburg (Kungliga Vetenskaps- och Vitterhets-Samhället i Göteborg).

References

Nobuhiko Akamatsu. 2005. Effects of second language reading proficiency and first language orthography on second language word recognition. *Second language writing systems*, pages 238–259.

David Alfter. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*. Ph.D. thesis, University of Gothenburg, Sweden.

David Alfter. 2024. Out-of-the-box graded vocabulary lists with generative language models: Fact or fiction? In *Proceedings of the 13th Workshop on NLP for Computer Assisted Language Learning*.

David Alfter, Yuri Bizzoni, Anders Agebjörn, Elena Volodina, and Ildikó Pilán. 2016. From Distributions to Labels: A Lexical Proficiency Analysis using Learner Corpora. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016*, 130, pages 1–7. Linköping University Electronic Press.

David Alfter, Rémi Cardon, and Thomas François. 2022. A dictionary-based study of word sense difficulty. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 17–24.

David Alfter and Johannes Graën. 2019. Interconnecting lexical resources and word alignment: How do learners get on with particle verbs? In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 321–326.

David Alfter, Therese Lindström Tiedemann, and Elena Volodina. 2021. Crowdsourcing relative rankings of multi-word expressions: Experts versus non-experts. In *Northern European Journal of Language Technology, Volume 7.*

David Alfter and Elena Volodina. 2018a. Is the whole greater than the sum of its parts? A corpus-based pilot study of the lexical complexity in multi-word expressions. In *Proceedings of SLTC 2018, Stockholm, October 7-9, 2018.*

David Alfter and Elena Volodina. 2018b. Towards Single Word Lexical Complexity Prediction. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 79–88.

Drilon Avdiu, Vanessa Bui, Klára Ptacinová Klimci, et al. 2019. Predicting learner knowledge of individual words using machine learning. In *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019), September 30, Turku Finland*, 164, pages 1–9. Linköping University Electronic Press.

Jill Burstein, Nitin Madnani, John Sabatini, Dan McCaffrey, Kietha Biggers, and Kelsey Dreier. 2017. Generating Language Activities in Real-Time for English Learners using Language Muse. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pages 213–215. ACM.

Annette Capel. 2010. A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1(1):1–11.

Jin-Woo Chung, Hye-Jin Min, Joonyeob Kim, and Jong C Park. 2013. Enhancing readability of web documents by text augmentation for deaf people. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, pages 1–10.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Press Syndicate of the University of Cambridge.

Scott A Crossley and Danielle S McNamara. 2012. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2):115–135.

Scott A Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359.

Jan De Belder, Koen Deschacht, and Marie-Francine Moens. 2010. Lexical simplification. In *Proceedings of ITEC2010 : 1st International Conference on Interdisciplinary Research on Technology, Education and Communication*.

Louise Deléger and Pierre Zweigenbaum. 2009. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora (BUCC)*, pages 2–10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Siobhan Devlin. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2018. Mining Words in the Minds of Second Language Learners for Learner-specific Word Difficulty. *Journal of Information Processing*, 26:267–275.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 897–906.

Núria Gala, Thomas François, Delphine Bernhard, and Cédrick Fairon. 2014. Un modèle pour prédire la complexité lexicale et graduer les mots. In *TALN 2014*, pages 91–102.

Núria Gala, Thomas François, and Cédrick Fairon. 2013. Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *E-lexicography in the 21st century: thinking outside the paper., Tallin, Estonia*.

Sian Gooding, Ekaterina Kochmar, Advait Sarkar, and Alan Blackwell. 2019. Comparative judgments are more consistent than binary classification for labelling word complexity. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 208–214.

Sian Gooding, Ekaterina Kochmar, Seid Muhie Yimam, and Chris Biemann. 2021. Word complexity is in the eye of the beholder. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4449.

Clarence Green, Anthony Kong, Marc Brysbaert, and Kathleen Keogh. 2025. Crowdsourced and AI-generated Age of Acquisition (AoA) Norms for Vocabulary in Print: Extending the Kuperman et al.(2012) norms. Preprint.

Marcel Adam Just and Patricia Ann Carpenter. 1987. *The psychology of reading and language comprehension*. Allyn & Bacon.

Abdelhak Kelious, Mathieu Constant, and Christophe Coeur. 2024. Investigating strategies for lexical complexity prediction in a multilingual setting using generative language models and supervised approaches. In *Proceedings of the 13th Workshop on NLP for Computer Assisted Language Learning*.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior research methods*, 44:978–990.

Lynn M LoPucki. 2014. System and method for enhancing comprehension and readability of legal text. US Patent 8,794,972.

Tomas Mikolov and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

Ian SP Nation and ISP Nation. 2001. *Learning vocabulary in another language*, volume 10. Cambridge university press Cambridge.

Joakim Nivre, Marie-Cathrine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of LREC*.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.

Gustavo Paetzold and Lucia Specia. 2016a. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Gustavo Paetzold and Lucia Specia. 2016b. Sv000gg at semeval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 969–974.

Charles A Perfetti, Edward W Wlotko, and Lesley A Hart. 2005. Word learning and individual differences in word learning reflected in event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(6):1281.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge Enhanced Contextual Word Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Workshop on Speech and Language Technology in Education*.

Alice Pintard and Thomas François. 2020. Combining Expert Knowledge with Frequency Information to Infer CEFR Levels for Words. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 85–92.

Irina Rets and Jekaterina Rogaten. 2021. To simplify or not? Facilitating English L2 users' comprehension and processing of open educational resources in English using text simplification. *Journal of Computer Assisted Learning*, 37(3):705–717.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *51st annual meeting of the association for computational linguistics proceedings of the student research workshop*, pages 103–109.

Matthew Shardlow, Fernando Alva-Manchego, Riza Theresa Batista-Navarro, Stefan Bott, Saul Calderon-Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, and Anna Huelsing. 2024. The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex: A New Corpus for Lexical Complexity Predicition from Likert Scale Data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI*, page 57.

Matthew Shardlow, Richard Evans, Gustavo Paetzold, and Marcos Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2022. Predicting lexical complexity in english texts: the complex 2.0 dataset. *Language Resources and Evaluation*, 56(4):1153–1194.

Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355.

Gerard J Steen, Aletta G Dorst, Tina Krennmayr, Anna A Kaal, and J Berenike Herrmann. 2010. A method for linguistic metaphor identification.

Anaïs Tack, Thomas François, Piet Desmet, and Cédrick Fairon. 2018. NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 137–146.

Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016. Evaluating Lexical Simplification and Vocabulary Knowledge for Learners of French: Possibilities of Using the FLELex Resource. In *LREC*.

Paul H Thibodeau and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PloS one*, 6(2):e16782.

Kevin Yancey and Yves Lepage. 2018. Korean L2 Vocabulary Prediction: Can a Large Annotated Corpus be Used to Train Better Models for Predicting Unknown Words? In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, United States. Association for Computational Linguistics.

Jiaxin Yuan, Cunliang Kong, Chenhui Xie, Liner Yang, and Erhong Yang. 2022. COMPILING: A Benchmark Dataset for Chinese Complexity Controllable Definition Generation. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 921–931, Nanchang, China. Chinese Information Processing Society of China.